The Patent Office

PCT/GB00/00492

INVESTOR IN PEOPLE

GB00/492

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

REC'D 13 APR 2000
WIPO PCT

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

I also certify that the attached copy of the request for grant of a Patent (Form 1/77) bears an amendment, effected by this office, following a request by the applicant and agreed to by the Comptroller-General.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1(a) OR (b)

Signed

Dated 04 APR 2000

Patents Form 1/77

Patents Act 1977
(Rule 16)

The Patent Office

Request for grant of a patent

(See the notes on the back of this form. You can also get an explanatory leaflet from the Patent Office to help you fill in this form)

19 FEB 1999

The Patent Office

Cardiff Road
Newport
Gwent NP9 1RH

1. Your reference

2. Patent application number
(The Patent Office will fill in this part)
9903697.2

3. Full name, address and postcode of the or of each applicant *(underline all surnames)*

PATTERNCOMPUTING LTD.
33, ARGYLE STREET
SOUTH BANK
YORK
YO23 1DW
UNITED KINGDOM

Patents ADP number *(if you know it)* 7605645001

If the applicant is a corporate body, give the country/state of its incorporation

4. Title of the invention A COMPUTER-BASED METHOD FOR MATCHING PATTERNS

5. Name of your agent *(if you have one)*

"Address for service" in the United Kingdom to which all correspondence should be sent *(including the postcode)*

~~PATTERNCOMPUTING LTD.~~
~~33, ARGYLE STREET~~
~~SOUTH BANK~~
~~YORK~~
~~YO23 1DW~~

AB
F51/77 3/4/00

URQUHART-DYKES & LORD
TOWER HOUSE
MERRION WAY
LEEDS
GB
LS2 8PA

Patents ADP number *(if you know it)* 07306085001

6. If you are declaring priority from one or more earlier patent applications, give the country and the date of filing of the or of each of these earlier applications and *(if you know it)* the or each application number

Country	Priority application number *(if you know it)*	Date of filing *(day / month / year)*

7. If this application is divided or otherwise derived from an earlier UK application, give the number and the filing date of the earlier application

Number of earlier application	Date of filing *(day / month / year)*

8. Is a statement of inventorship and of right to grant of a patent required in support of this request? *(Answer 'Yes' if:*

a) any applicant named in part 3 is not an inventor, or
b) there is an inventor who is not named as an applicant, or
c) any named applicant is a corporate body.
See note (d))

YES

Patents Form 1/77

9. Enter the number of sheets for any of the following items you are filing with this form. Do not count copies of the same document

Continuation sheets of this form	
Description	6
Claim(s)	3
Abstract	
Drawing(s)	

10. If you are also filing any of the following, state how many against each item.

Priority documents	
Translations of priority documents	
Statement of inventorship and right to grant of a patent *(Patents Form 7/77)*	
Request for preliminary examination and search *(Patents Form 9/77)*	
Request for substantive examination *(Patents Form 10/77)*	
Any other documents *(please specify)*	

11. I/We request the grant of a patent on the basis of this application.

Signature M. Turner Date 17/2/1999

(DIRECTOR: PATTERNCOMPUTING LTD)

12. Name and daytime telephone number of person to contact in the United Kingdom

M. TURNER
01904 679267

Warning

After an application for a patent has been filed, the Comptroller of the Patent Office will consider whether publication or communication of the invention should be prohibited or restricted under Section 22 of the Patents Act 1977. You will be informed if it is necessary to prohibit or restrict your invention in this way. Furthermore, if you live in the United Kingdom, Section 23 of the Patents Act 1977 stops you from applying for a patent abroad without first getting written permission from the Patent Office unless an application has been filed at least 6 weeks beforehand in the United Kingdom for a patent for the same invention and either no direction prohibiting publication or communication has been given, or any such direction has been revoked.

Notes

a) If you need help to fill in this form or you have any questions, please contact the Patent Office on 0645 500505.

b) Write your answers in capital letters using black ink or you may type them.

c) If there is not enough space for all the relevant details on any part of this form, please continue on a separate sheet of paper and write "see continuation sheet" in the relevant part(s). Any continuation sheet should be attached to this form.

d) If you have answered 'Yes' Patents Form 7/77 will need to be filed.

e) Once you have filled in the form you must remember to sign and date it.

f) For details of the fee and ways to pay please contact the Patent Office.

A Computer-Based Method for Matching Patterns.

PatternComputing Ltd,
33, Argyle Street,
Southbank,
York.
Company No: 3604291
Tel: 01904 679267

February 17, 1999

Description

Field of Invention

This is an invention in the area of pattern recognition. It is a general method for matching patterns held in computer memory, where a pattern is represented by a set of spatially or topologically arranged nodes each with an associated measurement vector.

Description of Current Techniques

There are a multitude of matching techniques. These may be split into two broad categories: gradient-based methods and exhaustive search. Examples of the former include gradient descent, simulated annealing, relaxation labeling, neural networks and genetic algorithms. All of these work by taking just a few initial best guess match solutions and refining them in order to obtain better solutions. The second category is exhaustive search. Here, a large number of match solutions are examined by sampling the solution space, and the best chosen. The forerunner of exhaustive search techniques is the fast access method called geometric hashing.

Problems with Current Methods

There are problems associated with both of the above categories. In short they are slow and give poor performance on non-trivial matching problems. There are reasons for this poor performance. Gradient-based methods depend critically on obtaining a good initial match, but this is obviously not possible in general since

having a good match is the aim in the first place. On the other hand exhaustive search methods are dependent on the resolution with which the solution space is searched. For matching the space is exponential in the number of nodes, making it very difficult to find a good solution in a reasonable time.

Overview of New Solution

We propose a new approach to matching which is fast and gives good performance. The approach stems from a new philosophy to pattern recognition based upon four key conditions:

- Condition 1

  Matching is formulated as one of finding the best set of transformations between the nodes in two patterns.

- Condition 2

  Calculations are underpinned by Bayesian probability theory.

- Condition 3

  The method is holistic in that it requires that all possible solutions must be examined.

- Condition 4

  Processing is resource-driven such that the calculations that can be performed are constrained by the memory available and the speed of operations required, as defined by the operator.

Conditions 3 and 4 lead to a conundrum: how to look at an exponential number of solutions quickly and efficiently. This is achieved by collecting solutions together into a small number of groups, and assessing each group in turn. There are a number of estimates that may be made on a group, but a strategy that is effective and consistent with condition 4 since we can trade-off speed for accuracy is to obtain upper and lower bound scores (probabilities) for any solution contained therein. Given these bounds the strategy to take is: eliminate groups of solutions if their upper bound falls below the highest lower bound. This guarantees that the optimal solution will be retained. By repeating this operation we can hone in on interesting regions of the solution space by excluding sub-optimal solutions. The remaining solutions may be re-examined in increasing detail as processing proceeds and as condition 4 allows. The process terminates when all upper bounds exceed the lower bound threshold. At this point the lower bound may be heuristically increased to re-start the elimination process, or alternatively the remaining transformations may be recorded and processed in some way.

Detailed Description of New Solution

Consider a pattern labelled by a set of N nodes. The nodes have an associated set of measurement vectors, $x = \{x_1, \ldots, x_N\}$.

Suppose that the set of transformations for the nodes is denoted by $w = \{w_i, \ldots, w_N\}$, say. From condition 1 the aim is to find the best global solution, i.e., the best set of transformations from the nodes in this pattern to a second pattern, where, from conditions 2 and 3 we adopt an holistic, probability theory approach, requiring:

$$w = \arg\max_{\hat{\omega} \in W} P(\omega = \hat{\omega} \mid x) \tag{1}$$

where W is the space of possible solutions. We do not realise this aim by directly, i.e., by actively searching for and refining solutions within W, this being the approach of existing gradient-based or exhaustive search techniques. Rather, we do so indirectly, by eliminating bad solutions from W. In doing so we implicitly examine all of the solution space, as required by condition 3, as follows.

We begin by grouping solutions together since examining each individual solution in isolation would be computationally intractible in general, thereby breaking condition 4. Consider all solutions that contain the individual transformation $w_i = \alpha$, say. The lowest upper bound on any one of these solutions is such that:

$$U(w_i = \alpha) = \max_{\tilde{w} \in \tilde{W}} P(w_i = \alpha, \tilde{w} \mid x) \tag{2}$$

where $\tilde{w}$ denotes the solutions on all nodes excluding that under consideration, and $\tilde{W}$ is the space of possible solutions for this set.

Now any group of solutions whose upper bound probability is below some known lower bound value, L, say, of interest cannot contain the optimum solution. Therefore, we can eliminate these groups from consideration. Therefore the rule at some iteration time n is:

eliminate any solution containing the transformation $w_i = \alpha$ *if*

$$U^{(n)}(w_i = \alpha) < L^{(n)} \tag{3}$$

This is the basis of the method: an upper bound on the probability of a group of solutions can be computed and compared against a lower bound threshold. If the upper bound falls below the threshold the group can be eliminated.

The computation of the upper bound has not yet been defined, and in general may be compuationally expensive, thereby breaking condition 4. The solution is to identify quantities of the form $G^{(n)}(w_i = \alpha)$ such that $G^{(n)}(w_i = \alpha) \geq U^{(n)}(w_i = \alpha)$ which can be computed in a given time. The elimination rule then becomes:

eliminate any solution containing the transformation $w_i = \alpha$ *if*

$$U^{(n)}(w_i = \alpha) \leq G^{(n)}(w_i = \alpha) < L^{(n)} \tag{4}$$

$G^{(n)}$ is evaluated by combining Bayesian probability theory with rules of inequality. Its form may change over the iterative cycles in order to accommodate condition 4. For example, at the onset of processing $G^{(n)}$ may be coarsely and quickly evaluated, but provided it obeys $G^{(n)} \geq U^{(n)}$ then only bad transformations will be eliminated. Towards the end of processing when only a few solutions remain, a more sophisticated and computatioanlly intensive means of computing G may be employed, such that $G^{(n)} \approx U^{(n)}$, provided condition 4 is not violated.

Processing will continue until no solutions fall below the threshold. At any time processing may be re-started by heuristically increasing the threshold, or alternatively, the remaining transformations may be recorded and processed in some manner.

An Example: Matching in Chemical Databases

An example use of the method is retrieval of bio-active compounds from chemical databases by using one or more query or lead compounds a cue. The starting point is to represent query and database compounds as patterns, each identified by a set of spatially or topologically arranged nodes, each node having an associated measurement vector.

We can develop the upper bound quantities. By applying Bayes' rule (2) becomes

$$U(w_i = \alpha) = \max_{\tilde{w} \in \tilde{W}} p(x \mid w_i = \alpha, \tilde{w}) P(w_i = \alpha, \tilde{w}) / p(x) \tag{5}$$

Making the non-restrictive assumption that the measurement vectors are independent when conditioned on the transformations then this becomes

$$U(w_i = \alpha) = p(x_i \mid w_i = \alpha) P(w_i = \alpha) \max_{\tilde{w} \in \tilde{W}} \{ \prod_{j \neq i} p(x_j \mid w_j) \} P(\tilde{w} \mid w_i = \alpha) / p(x) \tag{6}$$

Now introduce an inequality to reduce computational complexity. An option is $\max_{a \in A, b \in B} P(a, b) \leq max_{a \in A} P(a) \max_{b \in B} P(b)$ which gives

$$\begin{aligned} U(w_i = \alpha) \leq \\ p(x_i \mid w_i = \alpha) P(w_i = \alpha) \{ \textstyle\prod_{j \neq i} \max_{\beta \in W_j} p(x_j \mid w_j = \beta) \\ P(w_j = \beta \mid w_i = \alpha) \} / p(x) \end{aligned} \tag{7}$$

where W_j is the set of possible transformations for node j, and which reduces the complexity of the upper bound calculation from exponential to $O(N^2)$. Alternative inequalities could be applied here leading to increases or decreases in complexity, as required.

Now a current possibility can be eliminated as sub-optimal if its upper bound as computed in (5) falls below some determined lower bound threshold. For

example, if we have identified a solution and its probability at some time we can use this to set the lower bound. This gives a matching-by-exclusion algorithm in which regions of the solution space are iteratively pruned away. As a possible transformation on a node is eliminated at one time, so this affects the support computed for possibilities on other nodes at the next iteration.

The algorithm can be applied to all candidate transformations at all nodes, synchronously or asynchronously, and can be expressed as:

eliminate the transformation $w_i = \alpha$ ***from the list*** $W_i^{(n+1)}$ ***if***

$$p(x_i \mid w_i = \alpha)P(w_i = \alpha)\{\prod_{j \neq i} \max_{\beta \in W_j^{(n)}} p(x_j \mid w_j = \beta)P(w_j = \beta \mid w_i = \alpha)\} < \lambda_i^{(n)} \tag{8}$$

where $\lambda_i^{(n)}$ is the threshold value, and n is the time index.

Suppose that we take logarithms. The elimination rule then becomes

eliminate the transformation $w_i = \alpha$ ***from the list*** $W_i^{(n+1)}$ ***if***

$$S^{(n)}(w_i = \alpha) < \log \lambda_i^{(n)} \tag{9}$$

where $S^{(n)}(w_i = \alpha)$ counts the number of nodes that may be consistent with the assignment at node i:

$$\begin{aligned} S^{(n)}(w_i = \alpha) = \\ \log(p(x_i \mid w_i = \alpha)P(w_i = \alpha)) + \\ \textstyle\sum_{j \neq i} \max_{\beta \in W_j^{(n)}} \log(p(x_j \mid w_j = \beta)P(w_j = \beta \mid w_i = \alpha)) \end{aligned} \tag{10}$$

Application of the method in requires models for the distributions and priors in (10). For the application of compound matching one alternative is rectilinear distributions with zero height away from their centre. In this case the support for an individual transformation is:

$$S^{(n)}(w_i = \alpha) = k \sum_{j \neq i} \max_{\beta \in W_j^{(n)}} h(w_i = \alpha, w_j = \beta) \tag{11}$$

for $n > 0$, where k is a constant and where all solutions not compatible with the data have been eliminated at the onset. Here $h(w_i = \alpha, w_j = \beta)$ is a binary comptibility measure, simply stating if the transformation α on node i is compatible with the solution β on node j at time n.

The procedure can combine the algorithm in (9) with geometric hashing []. It involves a storage stage in which database compounds are encoded in a hash table, and a recall stage in which a query compound is used to access the table, and plausible transformations are examined. Finally, a clustering stage may be added to refine remaining solutions.

Storage

The following steps are taken in storage for each database compound:

- Generate the database compound nodes, and their measurement vectors to include node position and normal
- Generate a frame for each point using the centroid-position-normal triplet
- Align this frame to the world frame and store the compound in a hash table as compound-node-transformation triplets

Recall

The following steps are taken in recall:

- Generate the query compound to define the object nodes, their positions and normals
- Generate a frame for each node using the centroid-position-normal triplet
- Align this frame to the world frame and access the hash table, assigning accessed transformations to each node
- Convert the transformation matrices to rotation parameters and store in a hash table
- Use the matching-by-elimination procedure in (9) to eliminate implausible rotation solutions
- Cluster the remaining solutions and obtain a similarity index score for each by overlaying compounds

Modifications

Modifications to description above occur at the level of modeling. This may either be alterations to the form of the distributions assumed or to the measurement features employed. For example, in the molecule matching we have used rectilinear distributions but in this and other examples Gaussian distributions, say, may be appropriate and, for example, curvature information may have been employed.

Under different models and using different measurements there are a number of application areas for the technique:

- Medical image analysis
- Visual inspection and control
- DNA and protein sequence matching
- Financial prediction

A Computer-Based Method for Matching Patterns.

PatternComputing Ltd,
33, Argyle Street,
Southbank,
York.
Company No: 3604291
Tel: 01904 679267

February 17, 1999

Claims

- 1. A computer system of one or more processors for matching patterns by an examinination of all possible transformations between patterns, comprising:
  - a database of one or more patterns where each pattern is identified by a set of spatially or topologically arranged nodes, where each node has an associated measurement vector, the database being stored in one or more memories that are accessible to the processors;
  - a query pattern identified by a set of spatially or topologically arranged nodes, where each node has an associated measurement vector and an associated memory to store a set of possible transformations;
  - a method A for initialising the set of possible transformations on each node in the query pattern;
  - an iterative process for eliminating possible transformations, where each iteration comprises,

    a method B for computing an upper bound probability, U, on all solutions that contain a particular transformation in the set at a particular node;

    a method C for computing a lower bound probability, L;

    a thresholder for eliminating from the sets at each node transformations for which $U < L$.

    a halting procedure for stopping the iterative process.
  - a recorder for recording the transformations that remain.

- 2. A system, as in claim 1, where method B and method C may be designed or altered at any time so as to either increase or decrease the the time taken for processing to terminate, as required by the system operator, including the possibility of designing or altering method B and method C to be of constant complexity with the effect that the complexity of processing is then $O(N^2)$, where N is the number of nodes in the database patterns.

- 3. A system, as in claim 2, where further, method A, method B and method C are stipulated using Bayesian probability theory.

- 4. A system, as in claim 3, capable of accessing one or more objects from a database using a query of one or more objects, by a process of:

  - representing each object in the database as a pattern where each pattern is identified by a set of spatially or topologically arranged nodes, where each node has an associated measurement vector, the database being stored in one or more memories that are accessible to the processors;
  - selecting a query object from the query set;
  - representing the query object as a query pattern identified by a set of spatially or topologically arranged nodes, where each node has an associated measurement vector and an associated memory to store a set of possible transformations;
  - identifying the best transformations of the nodes in the query pattern as in claim 3, and optionally grouping these transformations by some averaging or clustering method, and subsequently assessing the probability or some similarity score of these remaining transformations;
  - recording those set of transformations and similarity scores.

- 5. A system, as in claim 4, capable of discovering novel bio-active compounds from a database using a query set of one or more lead compounds by representing database and query compounds as patterns, and for which method B and method C can be defined to be of variable complexity, including constant complexity for each node, leading to a complexity for processing of $O(N^2)$ for each compound-compound comparison, and providing for increases in the speed of analysis in excess of 2-3 orders of magnitude compared with existing techniques for discovering bio-active compounds.

- 6. A system, as in claim 4, capable of the identification of multi-dimensional objects at any degree of translation or rotation in image data, including the identification of advertising brands or logos in television or video data, for the purpose of registration or control, by representing the objects as patterns, and for which method B and method C can be defined to be of variable complexity, including constant complexity for each node, leading

to a complexity for processing of $O(N^2)$ and enabling the identification to be achieved in real time.